High End Computing Interagency Working Group (HECIWG) HEC File Systems and I/O Roadmaps

Rob Ross DOE/Office of Science ANL
Steve Poole DOE/Office of Science ORNL
Evan Felix DOE/Office of Science PNL
Bill Loewe DOE/NNSA LLNL
Lee Ward DOE/NNSA SNL
Gary Grider, John Bent, James Nunez DOE/NNSA LANL
Ellen Salmon NASA

Executive Summary

The need for immense and rapidly increasing scale in scientific computation drives the need for rapidly increasing scale in storage capability for scientific processing. Individual storage devices are rapidly getting denser while their bandwidth is not growing at the same pace. In the past several years, initial research into highly scalable file systems, high level Input/Output (I/O) libraries, and I/O middleware was conducted to provide some solutions to the problems that arise from massively parallel storage. To help plan for the research needs in the area of File Systems and I/O, the inter-government-agency published the document titled "HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the Fiscal 2005-2009 Time Frame" [Appendix C] which led the High End Computing Interagency Working Group (HECIWG) to designate this area as a national focus area starting in FY06. To collect a broader set of research needs in this area, the first HEC File Systems and I/O (FSIO) workshop was held in August 2005 in Grapevine, TX. Government agencies, top universities in the I/O area, and commercial entities that fund file systems and I/O research were invited to help the HEC determine the most needed research topics within this area. The HEC FSIO 2005 workshop report can be found at http://institute.lanl.gov/hec-fsio/docs/ . All presentation materials from all HEC FSIO workshops can be found at http://institute.lanl.gov/hec-fsio/workshops/
The workshop attendees helped

- catalog existing government funded and other relevant research in this area,
- list top research areas that need to be addressed in the coming years,
- determine where gaps and overlaps exist, and
- recommend the most pressing future short and long term research areas and needs necessary to help advice the HEC to ensure a well coordinated set of government funded research

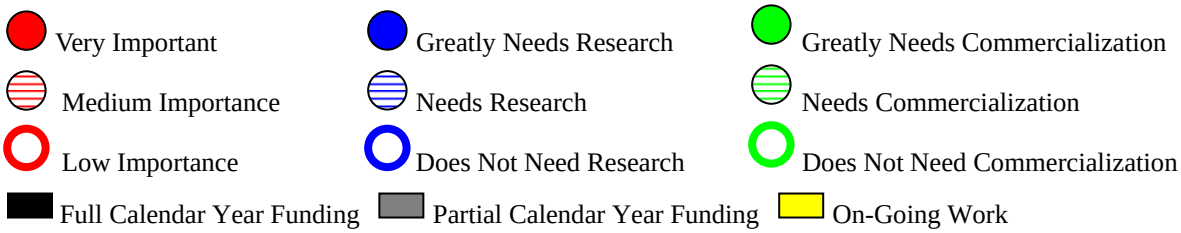
The recommended research topics are organized around these themes: metadata, measurement and understanding, quality of service, security, next-generation I/O architectures, communication and protocols, archive, and management and RAS. Additionally, university I/O center support in the forms of computing and simulation equipment availability, and availability of operational data to enable research, and HEC involvement in the educational process were called out as areas needing assistance.

Metadata Roadmap

Investigation into metadata issues is needed, especially in the areas of scalability, extensibility, access control, reliability, availability, and longevity for both file and archival systems. Additionally, consideration for very revolutionary ideas such as new approaches to name spaces and use of novel storage devices needs to be explored.

2007 Metadata Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Scaling	Bender/Farach-Colton	■	■	▒				 All existing work is evolutionary. What is lacking is revolutionary research; no fundamental solutions proposed.
	Leiserson	■	■	■				
	Maccabe/Schwann	■	■	▒				
	SciDAC - PDSI	■	■	■	■	■		
	HECEWG HPC Extensions	■	■	■	■	■	■	
	UCSC's Ceph	■	■	■	■	■	■	
	Lustre	■	■	■	■	■	■	
	ANL/CMU – Large Directory	■	■	■				
	PVFS	■	■	■	■	■	■	
Extensibility and Name Spaces	Bender/Farach-Colton	■	■	▒				 All existing work is evolutionary.
	Leiserson	■	■	■				
	Tosun	▒	■	■	▒			
	Wyckoff	■	■	▒				
	UCSC – LiFS/facets	■	■	■				
	ANL/CMU - MDFS		■	■				
	SciDAC PDSI	■	■	■	■	■		
File System/ Archive Metadata Integration	Lustre HSM	■	■	■	■			 Extended Attributes, although not standardized, could solve problem.
	UMN Lustre Archive	■	■					
Hybrid Devices Exploitation	*None*							 Research is being done, but no research focused on metadata
Data Transparency and Access Methods	*None*							 No research focused on metadata



Measurement and Understanding Roadmap

Research tools for measurement and understanding of parallel file system and end-to-end I/O performance are needed for advances in future file systems including evolutionary ideas such as layered performance measurement, benchmarking, tracing, and visualization of I/O related performance data. Also, more radical ideas like end-to-end modeling and simulation of I/O stacks and the use of virtual machines for large scale I/O simulation need to be explored.

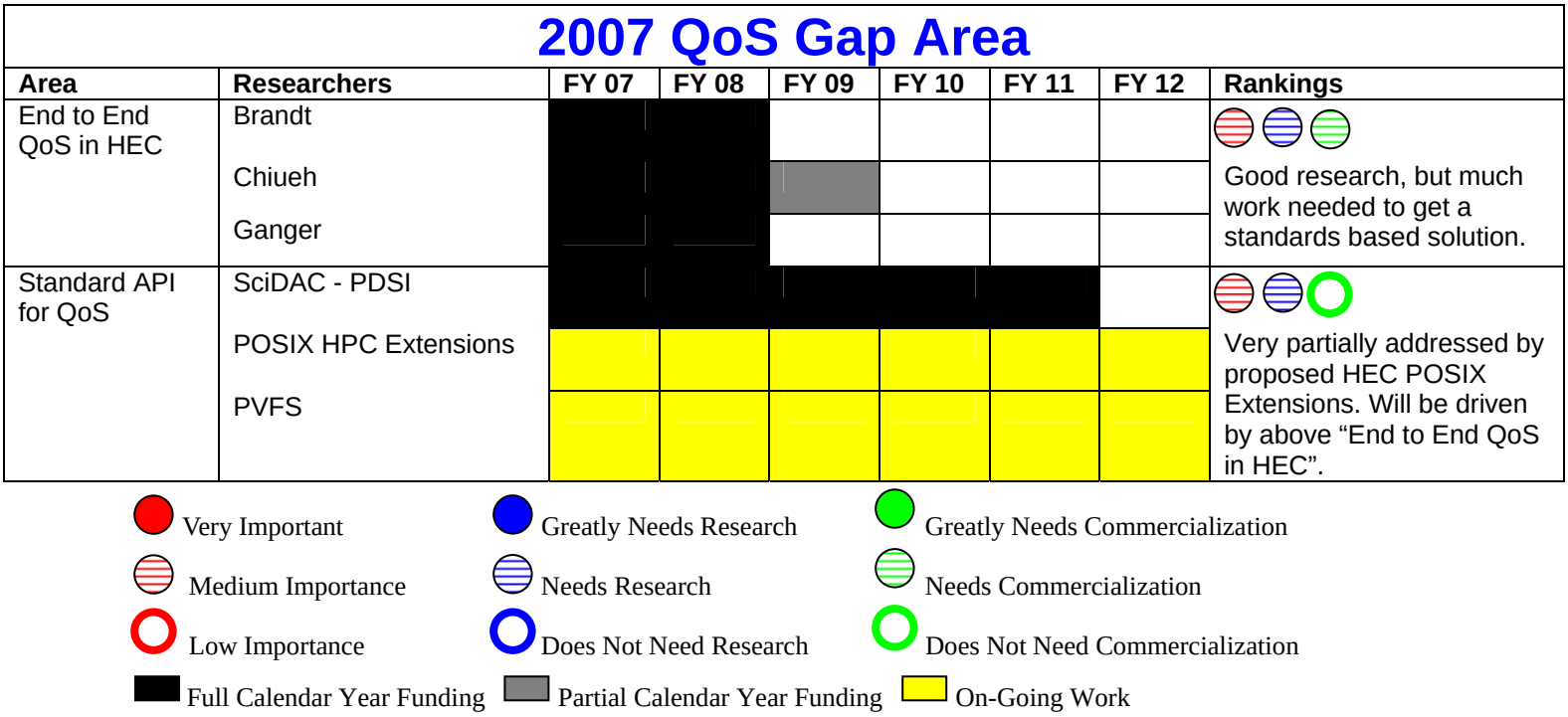
2007 Measurement and Understanding Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Understanding system workload in enterprise environment	Arpaci-Dusseau Reddy Zadok SciDAC - PDSI SciDAC - SDM							A comprehensive tool is nowhere in sight; problem is complex.
Standards for HEC I/O benchmarks	*None*							Low on agencies priorities; over simplifies problem and could drive vendors to incorrect solutions. Gap should really be replaced by release of traces, workload characterization, etc.
Testbeds for I/O Research	Ligon Thottethodi							Simulators are being developed. No real testbeds being built. This problem will only get worse over time, i.e. as systems get bigger.
Applying cutting edge visualization/ analysis tools to large scale I/O traces	Reddy Zadok							More traces are becoming available from Labs. Many opportunities to evaluate this research.

Very Important
Medium Importance
Low Importance
Greatly Needs Research
Needs Research
Does Not Need Research
Greatly Needs Commercialization
Needs Commercialization
Does Not Need Commercialization
Full Calendar Year Funding
Partial Calendar Year Funding
On-Going Work

Quality of Service Roadmap

Quality of service (QoS) can be defined as features of a storage architecture that allow a user or administrator to recommend policies for data movement during I/O operations. QoS is a ripe topic for research especially in the area of providing prioritized, deterministic performance in the face of multiple, complex, parallel applications running concurrently with other non-parallel workloads. More revolutionary ideas such as dynamically adaptive end-to-end QoS throughout the hardware and software I/O stack are highly desirable.



Next-generation I/O Architectures Roadmap

I/O stacks and architectures have been static for some time now forcing developers to adopt awkward solutions in order to achieve target I/O rates. There is great need for research into next-generation I/O architectures, including evolutionary concepts such as extending the POSIX I/O API standard to support archives in a more natural way, access awareness, and HEC/high concurrence. Studies into methods to deal with small, unaligned I/O and mixed-size I/O workloads as well as collaborative caching and impedance matching are also needed. Novel approaches to I/O and file systems also need to be explored including redistribution of intelligence, adaptive and reconfigurable I/O stacks, user space file systems, data-aware file systems, and the use of novel storage devices. This area may be well-served by delving into and applying the research from the modeling community.

2007 Next Generation I/O Architectures Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Understanding file system abstractions - File system architectures	Choudhary Dickens Maccabe/Schwan Reddy Shen Thain Wyckoff SciDAC – PDSI PNNL							Good work, but much of research is in infancy. A small portion ready for commercialization.
Understanding file system abstractions - naming and organization	Bender/Farach-Colton Thain Tosun Zhang/ Jiang SciDAC – SDM SciDAC - PDSI							Very hard problem. More researchers need to attack this problem.
Self-assembling, Self-reconfiguration, Self-healing storage components	Ganger Ligon Ma/Sivasubramaniam/ Zhou SciDAC - PDSI SciDAC - SDM							Good work being done, but it's a hard problem that will take more time to solve.
Architectures using 10^6 storage components	Ligon PNNL							Very little work being done here for a very near term problem. Simulators will/must play a role here
Hybrid architectures leveraging emerging storage technologies	Gao PNNL							Big potential reward, but very little work being done in the HPC area.

2007 Next Generation I/O Architectures Gap Area

Area	Researcher	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
HEC systems with multi-million way parallelism doing small I/O operations	Choudhary							Good initial research; needs to be moved into testing. More fundamental solutions being pondered including non-volatile solid state store.
	Dickens							
	Gao							
	FASTOS – I/O Forwarding							

Very Important
Greatly Needs Research
Greatly Needs Commercialization
Medium Importance
Needs Research
Needs Commercialization
Low Importance
Does Not Need Research
Does Not Need Commercialization
Full Calendar Year Funding
Partial Calendar Year Funding
On-Going Work

Communication and Protocols Roadmap

In the area of file system related communications and protocols, evolutionary items such as exploitation of Remote Direct Memory Access (RDMA), Object Based Secure Disk (OBSD) extensions, Network File System Version 4 (NFSv4) extensions, and parallel Network File System (pNFS) proof-of-concept implementations as well as more revolutionary exploration of server to server communications are needed.

2007 Communication and Protocols Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Active Networks	Chandy							Novel work being done, but not general enough.
Alternative I/O transport schemes	Sun							Most aspects are being addressed.
	Wyckoff							
	Lustre							
	pNFS							
Coherent Schemes	ANL/CMU							No consensus on how to do this correctly, but some solutions are in products.
	UCSC's Ceph							
	Lustre							
	Panasas							
	PVFS							

Very Important
Greatly Needs Research
Greatly Needs Commercialization
Medium Importance
Needs Research
Needs Commercialization
Low Importance
Does Not Need Research
Does Not Need Commercialization
Full Calendar Year Funding
Partial Calendar Year Funding
On-Going Work

Archive Roadmap

In the area of archive, the interfaces to the file systems and I/O stacks in HEC systems and long term care for the massive scale of an archive in the HEC environment are difficult areas needing more research than they have received before.

2007 Archive Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
API's/Standards for interface, searches, and attributes, staging etc.	Ma/Sivasubramaniam/ Zhou Tosun SciDAC – SDM SciDAC – PDSI							Current research is in terms of file systems, not archive. API merging with POSIX and API for searching lacking
Long term attribute driven security	Ma/Sivasubramaniam/ Zhou Odlyzko							Current research is in terms of file systems, not archive. Current researchers need data supporting proposed solutions usefulness
Long term data reliability and management	Arpaci-Dusseau Narasimhan							Need for commercialization is low because of other drivers, i.e. HIPPA and others will drive this. Redundancy techniques reasonably sufficient for archives
Metadata scaling	Bender/Farach-Colton Jiang/Zhu Leiserson Ganger Panasas Lustre ANL/CMU							Current research is in terms of file systems, not archive, but this work can be applied to archive. File system research will be more than fast enough for archive.
Policy driven management	*None*							Sarbanes-Oxley Act is solving this problem

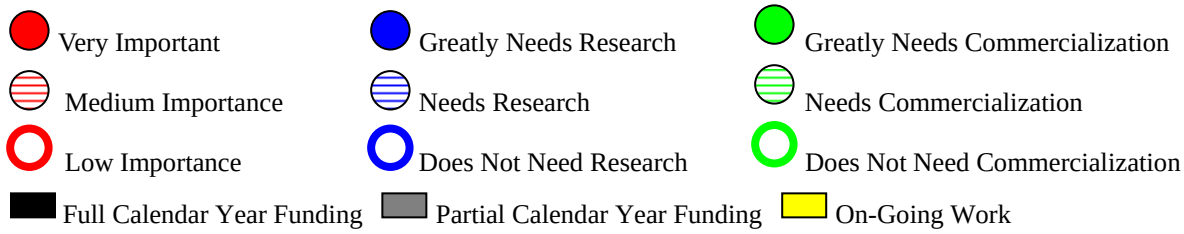
Very Important
Medium Importance
Low Importance
Greatly Needs Research
Needs Research
Does Not Need Research
Greatly Needs Commercialization
Needs Commercialization
Does Not Need Commercialization
Full Calendar Year Funding
Partial Calendar Year Funding
On-Going Work

Management and RAS Roadmap

In the area of management, reliability and availability at scale, management scaling, continuous versioning, and power management are all needed research topics. Additionally, more revolutionary ideas like autonomics, use of virtual machines, and novel devices exploitation need to be explored.

2007 Management and RAS Gap Area

Area	Researchers	FY 07	FY 08	FY 09	FY 10	FY 11	FY 12	Rankings
Automated problem analysis and modeling	Reddy							More researchers need to look at this problem.
Formal Failure analysis for storage systems	Arpachi-Dusseau							Good research done here. Will people use this work?
Improved Scalability	Ganger							More research is needed here. Testbed is probably needed for this work.
	Ligon							
Power Consumption and Efficiency	Qin							Industry is working on this problem. Storage is not a large consumer of energy at HEC sites.
Reliability	*None*							Industry is working on this problem



Security Roadmap

Aspects of security such as usability, long term key management, distributed authentication, and dealing with security overhead are all topics for research. There is also room for more difficult research topics such as novel new approaches to file system security including novel encryption end-to-end or otherwise that can be managed easily over time. The need for standardization of access control list mechanisms is also needed and investigation into a standard API for end-to-end encryption could be useful.

2007 Security Gap Area

Area	Researchers	CY 06	CY 07	CY 08	CY 09	CY 10	CY 11	Rankings
Long term key management	Odlyzko							Current researcher need data to validate designs
End-to-end encryption	Odlyzko							Current researcher need data to validate designs
Performance overhead and distributed scaling	Sivasubramaniam							Problem reasonably well understood, unclear if enough demand for product
Tracking of information flow, provenance, etc.	*None*							Industry will help some, but not in HPC context. Nothing to commercialize yet.
Ease of use, ease of management, quick recovery, ease of use API's	Sivasubramaniam							Current researchers need data to validate designs Nothing to commercialize yet.

Very Important
Medium Importance
Low Importance
Greatly Needs Research
Needs Research
Does Not Need Research
Greatly Needs Commercialization
Needs Commercialization
Does Not Need Commercialization
Full Calendar Year Funding
Partial Calendar Year Funding
On-Going Work

Assisting with Standards, Research and Education Roadmap

At the HEC FSIO 2005 workshop, there was a recognition that the HEC FSIO community should find ways of supporting students working in the general area of I/O as well as students working more specifically on I/O within HEC. Investment to support the research of these students was considered worthwhile both because they may provide important research while still in school as well as by cultivating these students such that they may continue to work on HEC I/O problems following their graduation and, with any luck, become the next generation of HEC I/O experts.

Over the past decade, the HEC community has had a role in the formation and adoption of various FSIO related standards. The most notable are the ANSI T10 1355D specification for Object Based Storage Devices (OBSD), the IETF NFSv4 standard including the new pNFS portion of the NFSv4.1 minor revision of the NFSv4 specification, and the newly formed Open Group HEC Extensions to the POSIX standards work has also been an outcome of HEC FSIO and the HEC I/O community work. Past years are status, future years are identified needs or desires

2007 Assisting with Standards, Research and Education					
Area	FY07	FY 08	FY 09	FY 10	FY 11
Standards: POSIX HEC ANSI OBSD IETF pNFS	 PDSI UM CITI patch pushing/maintenance Revamp of manual pages V2 nearing publication V 4.1 nearing pub Assistance in testing may be needed	 First Linux full patch set Some file system pilot test Initial products			
Community Building	HEC FSIO 2007 HEC presence at FAST and IEEE MSST	HEC FSIO 2008 HEC presence at FAST and IEEE MSST	HEC FSIO 2009 HEC presence at FAST and IEEE MSST	HEC FSIO 2010 HEC presence at FAST and IEEE MSST	HEC FSIO 2011 HEC presence at FAST and IEEE MSST
Equipment	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility	Incite and NSF Infra Need scale CS disruptive facility
Simulation Tools	Ligon PDSI Felix/Farber	Ligon PDSI Felix/Farber	Ligon PDSI Felix/Farber		
Education	LANL Institutes as one example PDSI	Other Institute like activities			
Research Data	Failure, usage, event data	Many more traces, FSSTATS, more disk failure data			

Conclusion

In the near future, sites will deploy supercomputers with hundreds of thousands processors routinely. Million-way parallelism is around the corner and, with it, bandwidth needs to storage will go from tens of gigabytes/sec to terabytes/sec. Online storage requirements to support work flows for efficient complex science will begin to approach the exabyte range. The ability to handle a more varied I/O workload ranging seven orders of magnitude in performance characteristics, extremely high metadata activities, and management of trillions of files will be required. Global or virtual enterprise wide-area sharing of data with flexible and effective security will be required. Current extreme-scale file system deployments already suffer from reliability and availability issues, including recovery times from corruption issues and rebuild times. As these extreme-scale deployments grow larger, these issues will only get worse. It will possibly be unthinkable for a site to run a file system check utility, yet it is almost a given that corruption issues will arise. Recovery times need to be reduced by orders of magnitude, and these types of tools need to be reliable, even though they may rarely be used. The number of storage devices needed in a single coordinated operation could be in the tens to hundreds of thousands, requiring integrity and reliability schemes that are far more scalable than available today. Management of enterprise-class global parallel file/storage systems will become increasingly difficult due to the number of elements involved, which will likely approach 100,000 spinning disks with widely varying workloads. The challenges of the future are formidable.